Decision Support System for Keyword Spotting Using Theory of Evidence

Wasiq Khan*, Rob Holton
School of Electrical Engineering & Computer Science, University of Bradford, BD7 1DP, United Kingdom.

* Corresponding author. Email: wkhan6@bradford.ac.uk
Manuscript submitted August 19, 2015; accepted January 20, 2016.
doi: 10.17706/ijcee.2016.8.1.22-30

Abstract: Keyword Spotting (KWS) in continuous speech is an emerging but challenging task that needs to deal with speech dynamics. Literature contains a variety of approaches for keyword spotting and spoken term detection. Most of these techniques are based on pattern based methods that are limited to a specific vocabulary of words with high computational cost for system training. A valuable contribution to the existing speaker dependent keyword spotting approaches is made in this paper by introducing a template matching based approach that exploits the Dempster's theory of Combined Evidence (CEv). The CEv plays a significant role as a Decision Support System (DSS) by integrating the keyword match/mismatch beliefs from multiple resources and providing a combined score (belief). Finally, performance of the proposed approach is compared with the existing keyword spotting techniques by using statistical analysis of the experimental results.

Key words: Spoken term detection, template matching, decision support system, similarity measure.

1. Introduction

In the template based speech modeling, recognition is performed by matching the test word (utterance) with the all stored templates of words and calculating the matching score based on acoustic features [1]. The Dynamic Time Warping (DTW) and Vector Quantization (VQ) based speech recognition are the best examples of such systems. A KWS system is based on the partial information extraction (keyword) from a continuous speech signal. Despite of the fact that research has been conducted in the area of KWS since forty years, yet the formulation of the KWS has not been well established [2]. The related research work can be summarized into three main categories that include Query-by-Example (QbyE) methods, keyword/filler methods, and Large Vocabulary Continuous Speech Recognition (LVCSR) methods. Among the aforementioned approaches, QbyE is the most relevant to the approach proposed in this research study. Literature consists of a number of KWS approaches in relation to QbyE that use some sort of variations in DTW [3]-[6].

Over the past decade, most of the related research is focused on novelty of the template representation methods [7]-[10]. An unsupervised spoken term detection using acoustic segment model is presented by [10]. The aim of the study was to measure the QbyE performance using acoustic segmentation model based posteriorgrams and traditional Gaussian mixture model posteriorgrams. The acoustic segment models are the unsupervised Hidden Markov Models (HMM) of non-transcribed speech data. A segmented DTW is applied for the query and test utterance matching and the location of the query utterance are identified.

The Fisher and the TIMIT dataset are used for the experimentation purpose. System performance is measured using the standard binary classification method. Despite of the fact that this approach does not uses the speech transcript for the supervised training, it uses the HMMs for posteriorgrams representation that may take huge amount of computational cost.

Spoken term detection in speech for QbyE approach is introduced in [8] for a limited or no in-domain training data. The keyword and template speeches are represented by phonetic posteriorgrams obtained from a phonetic recognition system. The measured posteriorgrams are forwarded to a constrained DTW that measures the warping distance and the position with minimum warping distance is identified as desired keyword. Advantage of this approach is the language independence because the transcript data is not used. However, the accuracy in terms of keyword detection may be a question that is improved in the proposed research work. Similarly, a keyword spotter presented in [11] is based on MFCC and energy of the speech signal as feature set. The system is analyzed using Hamming window with 36 milliseconds frame length. The VQ algorithm is used for the vector training and codebook generation. Each codebook represents the acoustic features of the signal. The HMMs were used for the probability assignment for observation given a word. This system is based on VQ and HMM that need a high computational time for training learning the models. Moreover, HMM based recognizers depend upon probability assignment that may not be assigned correctly and may cause misrecognition.

Existing pattern matching based KWS approaches uses different versions of DTW to resolve the time warping issue. However, there are some challenges associated with the DTW approach that are needed to be considered. Firstly, the distance matrix of DTW grows exponentially with increasing length of the speech utterance. This issue can be resolved by applying some boundary constraints and pruning the search space [12]; however the pruning process may lose important information leading to a considerable sacrifice of the performance. Secondly, DTW uses the information of a single distance metric to make the match/mismatch decision about of a query utterance. As the DTW model doesn't use the transcribed data for training, single information resource may not be reliable. In this paper, a robust method for KWS is proposed that is based on acoustic features and resolves the aforementioned challenges associated with DTW. Very first time, CEv is deployed as DSS that fuses the beliefs from multiple distance metrics and provides a combined matching score that is more reliable.

Section 2 of this paper presents the sequential flowchart followed by the formulation of Dempster's theory of evidence. Performance evaluation and simulation settings are presented in Section 3. A detailed discussion on KWS results is presented using binary classifications metrics.

2. Proposed Method

A KWS can be considered as a sub-part of the automated speech recognition which aims to extract the partial information from speech signal in the form of a query utterance (keyword). Time warping effects due to dynamic length of spoken words and existence of the silence segments are the most significant challenges to be resolved. A number of methodologies are amalgamated sequentially to resolve the speech dynamics and time warping issues in the proposed KWS approach. Fig. 1 show the sequential flow of the all processes and detail of each component is addressed in the following section.

2.1. Pre-processing and Feature Extraction

Most of the time, speech signal consists of silence parts and background noise that may be a major cause of mismatching or misidentification. Pre-treatment is a process to enhance the input speech signals in terms of sample rate, silence removal, and background noise reduction. There are a number of techniques in the literature that uses time and frequency domain features (i.e. energy, zero cross rate, spectral centroid) to remove the silence part of speech signal [13], [14]. In the proposed KWS approach, an efficient silence

removal method is used that was introduced in our previous work [15]. This approach deploys a pitch tracking method proposed in [16] to estimate the fundamental frequency using multiple information resources. The enhanced speech signals are then forwarded for the feature extraction process to be expressed as a sequence of feature vectors that may provide sufficient information to represent the speech utterance. The MFCCs are the most dominant and distinguishing features of human speech that have been successfully used in the literature [15], [17]. Rather than extracting the traditional MFCC features from the speech signal, wavelet based features are also extracted. The combined feature set of MFCC's mean values and Wavelet Energy (WE) measurements empowers the KWS performance that is discussed later in the performance section.

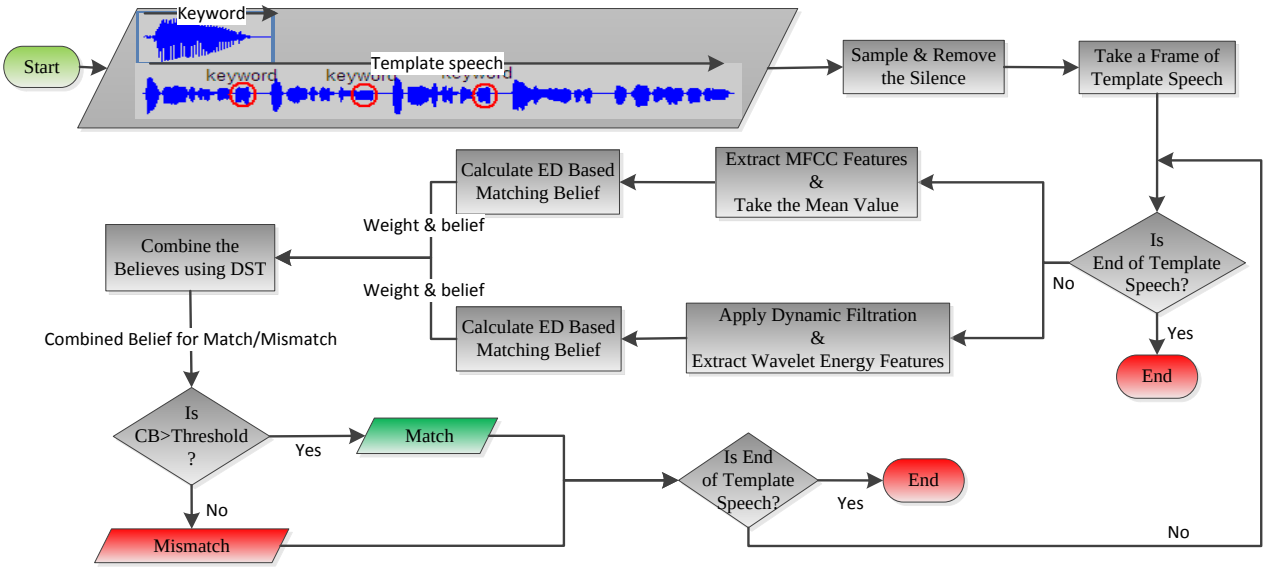

Fig. 1. Processing flow of the proposed method.

A major advantage of wavelet transform is the simultaneous representation of time and frequency analysis which helps to filter out the noisy segments from the spectrum as well as from the time domain speech signal. Wavelet decomposition analyses the function at various levels of resolution and provides a simultaneous time-frequency representation of input speech signal. This representation empowers the wavelet's superiority because of its efficiency for localizing the frequency in time domain along with the correlation matrix as a third dimension. Approximation and detailed coefficients possessing the magnitude values below the threshold are filtered out while remaining coefficients are integrated together and forwarded for further processing. Energy in a frequency level is measured by integrating the intensity magnitudes over time. After calculating the normalized energy for each frequency band, a threshold value (0.7) is applied to each frequency band to filter out the unnecessary scales. The extracted MFFC and WE features for both; keyword and template frame are first normalized and then forwarded to a similarity measure. Euclidean distance is used as a similarity measure. As Euclidean distance provides dissimilarity score, fewer score means more similar. Because of the normalized data distribution, the similarity measurement values are within the ranges of 0 and 1 for each pair of keyword and template speech frame. These measurements are used as the match/mismatch beliefs that are forwarded to DSS along with the pre-set weights.

2.2. Formulation of the Beliefs Combination

The Dempster-Shafer Theory (DST) is considered as a generalization to the Bayesian theory in such a way that it can handle the degree of ignorance. In case of certain information, there are a number of fusion

methods that can provide a combined belief. However, most of these approaches are unable to handle the degree of ignorance. The DST provides the best estimate of the degree of belief by combination of evidences/ believes from multiple resources. A detailed study on DST advantages, disadvantages, criticism, and its application areas is presented in [18].

2.2.1. Define the basic attributes

Let's E={*mel*, *e*} represents the set of basic attributes for the proposed KWS where '*mel*' and '*e*' are the belief resources. In our case, MFCC and WE are such resources. The relative weights for the basic attributes are pre-set by offline experiments such that $0 \leq \omega_i \leq 1$ and it fulfils (1).

$$\sum_{i=1}^{L} \omega_i = 1 \tag{1}$$

where; 'L=2' represent the number of attributes that are (*mel*, *e*) for KWS. The distinctive evaluation grades are defined as set of two entities, i.e. H= {match, mis_match}. For each attribute in 'E' and evaluation grade 'H', a degree of belief β_n is assigned. The degree of belief denotes the source's level of confidence when assessing the level of fulfillment of a certain property.

2.2.2. Basic probability assignments for each basic attribute

Let $m_{n,i}$ be a basic probability mass representing the degree to which the i^{th} basic attribute. A hypothesis that the general attribute is assessed to the n^{th} evaluation grade H_n can be presented as:

$$m_{n,i} = \omega_i \beta_{n,i} \tag{2}$$

where 'n' are number of evaluation grades (match, mis_match). The remaining probability mass $m_{H,i}$ unassigned to each basic attribute is calculated as:

$$m_{H,i} = 1 - \sum_{n=1}^{N} m_{n,i} = 1 - \omega_i \sum_{n=1}^{2} \beta_{n,i} \tag{3}$$

where, 'N=2' are the total number of evaluation grades. The remaining probability mass is further decomposed into $\bar{m}_{H,i}$ and $\tilde{m}_{H,i}$ as:

$$\bar{m}_{H,i} = 1 - \omega_i \tag{4}$$

$$\tilde{m}_{H,i} = \omega_i \left(1 - \sum_{n=1}^{2} \beta_{n,i} \right) \tag{5}$$

With

$$m_{H,i} = \bar{m}_{H,i} + \tilde{m}_{H,i} \tag{6}$$

Equation (4) measures the degree to which final attributes have not yet been assessed to individual grades due to the relative importance of basic attributes after their aggregation. Equation (5) measures the degree to which final attributes cannot be assessed to individual grades due to the incomplete assessments for basic attributes.

2.2.3. Combined probability assignments

In this step, the probability mass of the basic attributes E={*mel*, *e*} are aggregated to form a single assessment for keyword match/mismatch. The combined probability masses can be generated using the following set of recursive evidence reasoning equations:

$$\begin{aligned} &\{H_n\}: \\ &m_{n,i+1} = K_{i+1}[m_{n,i}.m_{n,i+1} + m_{H,i}.m_{n,i+1} + m_{n,i}.m_{H,i+1}] \\ &n = 1,\ldots,N \end{aligned} \tag{7}$$

where i={1,..., L-1}, L=2 is the number of basic attributes, and 'N=2' are the total number of evaluation grades. In (7), $m_{n,1}.m_{n,2}$ measures the degree of both attributes {*mel*, *e*} supporting the general attribute of keyword match to be assessed to H_n. The term $m_{n,1}.m_{H,2}$ measures the degree of only 1st attribute {*mel*} supporting keyword match to be assessed to H_n. The term $m_{H,1}.m_{n,2}$ measures the degree of only 2nd attribute {*e*} supporting final belief to be assessed to H_n.

$$\{H\}: m_{H,i} = \bar{m}_{H,i} + \tilde{m}_{H,i} \tag{8}$$

$$\tilde{m}_{H,i+1} = K_{i+1}[\tilde{m}_{H,i}.\tilde{m}_{H,i+1} + \bar{m}_{H,i}.\tilde{m}_{H,i+1} + \bar{m}_{H,i+1}.\tilde{m}_{H,i}] \tag{9}$$

$$\bar{m}_{H,i+1} = K_{i+1}[\bar{m}_{H,i}.\bar{m}_{H,i+1}] \tag{10}$$

$$K_{i+1} = \left[1 - \sum_{t=1}^{N=2} \sum_{\substack{j=1 \\ j \neq t}}^{N=2} m_{t,i}.\, m_{j,i+1}\right]^{-1} \tag{11}$$

where, i={1,..., L-1}. In (9), $\tilde{m}_{H,1}.\tilde{m}_{H,2}$ measures the degree to which final attribute cannot be assessed to any individual grades {match, mis_match} due to the incomplete assessments for both attributes {*mel*, *e*}. Term $\bar{m}_{H,1}.\tilde{m}_{H,2}$ measures the degree to which final attributes cannot be assessed due to the incomplete assessments for {*mel*} only. In (10), $\bar{m}_{H,1}.\bar{m}_{H,2}$ measures the degree to which final attribute has not yet been assessed to individual grades due to the relative importance of {*mel*} and {*e*} after {*mel*} and {*e*} have been aggregated. The normalization factor 'K' is used to normalize m_n, m_H such that $\sum_{n=1}^{N=2} m_n + m_H = 1$.

2.2.4. Calculation of the combined degree of belief

Let β_n denote the combined degree of belief that the KWS assessed to the grade H_n, which is generated by combining the assessments for all the associated basic attributes E={*mel*, *e*}, then β_n is calculated by:

$$\{H_n\}: \beta_n = \frac{m_{n,L}}{1 - \bar{m}_{H,L}} \qquad n = 1,..,N \tag{12}$$

$$\{H\}: \beta_H = \frac{\tilde{m}_{H,L}}{1 - \bar{m}_{H,L}} \tag{13}$$

Above equation for β_H measures the belief that is left unassigned during the assessments.

3. Performance Evaluation and Experimental Setup

A number of metrics have been used in the literature to evaluate the performance of KWS approaches. However, the most relevant are the gold standards used for the binary classification [19]. This is because the output of KWS is in the binary form (match or mismatch). To conduct a case study in this research, a SENNHEISER e935 is used for speech recordings which is a vocal dynamic microphone consisting a built in noise filter. Speech is recorded at a sampling frequency of 8 KHZ. To conduct a case study, we have recorded a speech dataset by 30 speakers (17 male, 13 female) that consists of connected words in the form of digits (5 recordings for each digit by each speaker), short phrases of up to 10 seconds (5 sentences by each speaker) and long phrases of up to 20 seconds (5 paragraph bay each speaker). For the long speech phrase spotting experiments; a speech corpus from American Rhetoric's (top 100 speeches) [20] is used. It is based on hours of speeches recorded by different people on different topics. Because of the template frames overlapping, a mismatch tolerance of single frame size is set throughout the experiment conduction. Individual performances of the proposed combined evidence (CEv) based KWS approach is compared with the existing constrained DTW and segmented-DTW based KWS approaches.

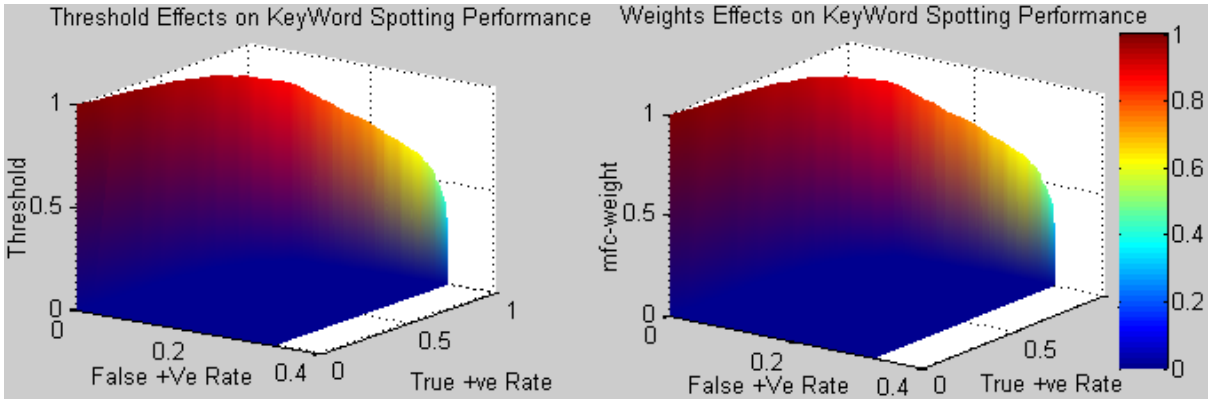

Fig. 2. Setting the threshold value and weights for belief resources.

An important factor for the proposed KWS performance is the threshold value that specifies the decision boundary for the keyword to be considered as a match or mismatch. The trade-off between sensitivity and specificity depends upon the threshold value change. Increase in the threshold value will increase the specificity and vice versa. However, the threshold value can be set with respect to the application area. For example, KWS used for the intelligent agencies may assign more importance to sensitivity to maximize the true positives by reducing the threshold. In such scenarios, the objective is to identify the desired keyword (e.g. blast, terror) that exists in a speech recording. To set a threshold value for match/mismatch decision boundary, the ROC curve is achieved by varying threshold from 0 to 1 with a lag of 0.01 as shown in Fig. 2. It is observed that the best value in ROC curve is achieved with a threshold value of 0.85 (85%). It means that the template frame will be rejected if its matching belief with the keyword is less than 85%. As there is a trade-off between sensitivity and specificity, threshold value is chosen while considering both metrics. As defined in (2, 3) that the basic probabilities in DST depend upon the weights assigned to the basic attributes E={*mel*, *e*}. Experiments are conducted by setting continuously varying weights for both attributes from 0 to 1 with a lag of 0.01. The ROC curve is achieved (Fig. 2) for 100 values of weights between 0 and 1. It is observed that the best performance in terms of FPR and TPR is achieved at w_{mfcc} =0.75, w_{wav} = 0.25. This implies that the best performance is achieved by assigning more weight to matching belief of MFCC based features.

Table 1. Performance Comparison of Proposed Method and DTW

Performance Metrics		CEv	Wav	MFCC	DTW
Sensitivity		0.97	0.92	0.95	0.83
Specificity		0.93	0.86	0.92	0.90
Accuracy		0.92	0.86	0.93	0.88
1/LR+		0.06	0.14	0.07	0.17
LR-		0.03	0.07	0.05	0.43
True +Ve Rate		0.97	0.93	0.96	0.29
Type I Error	**μ**	0.01	0.02	0.01	0.01
	σ	0.01	0.02	0.01	0.01
Type II Error	**μ**	0.01	0.01	0.03	0.26
	σ	0.02	0.03	0.10	0.27

It can easily be observed that the sensitivity of CEv is greater than the individual values of MFCCs and wavelets by a factor of 2% and 5% respectively. This implies that the deployment of DST increases the KWS as well as it empowers the performance in terms of decision making. The likelihood ratios (LR+, LR-) are considered one of the best metrics to measure the diagnostic accuracy. In terms of KWS, LR presents the probability of a test with keyword match divided by the probability of the same test with keyword mismatch. Larger LR+ consist more information than smaller LR+ whereas smaller LR- consists more information than larger LR-. To simplify the LR values, a relative magnitude is considered by taking the reciprocal of LR+. It is analysed from Table 1 that LR- for the CEv approaches to zero (0.03) as compared to 0.4 for DTW. This indicates the superiority of the proposed KWS approach over the existing DTW based KWS approaches.

Type I and Type II errors indicate the recognizer failure related to FP and FN respectively. These metrics have been represented in a number of ways in the related area that include mean square error and absolute errors as most common metrics. Table 1 demonstrates 'μ' (mean) and 'σ' standard deviation for both types of error for five different approaches while using the same dataset. As discussed before, in KWS related tasks, Type II error may have more importance as compared to Type I error because of the more emphasis on spotting a keyword. However, it may vary with respect to application area. It is observed in Table 1 that the 'μ' and 'σ' for Type II error are negligible (i.e. 0.006 and 0.019 respectively) in case of CEv based KWS as compared to DTW approach (0.25 and 0.27 respectively) which indicates the robustness of our approach. In addition to this, it can also be observed that the individual errors for MFCCs and wavelet based approaches are higher than the CEv approach that proves the significance of the proposed DSS for KWS task.

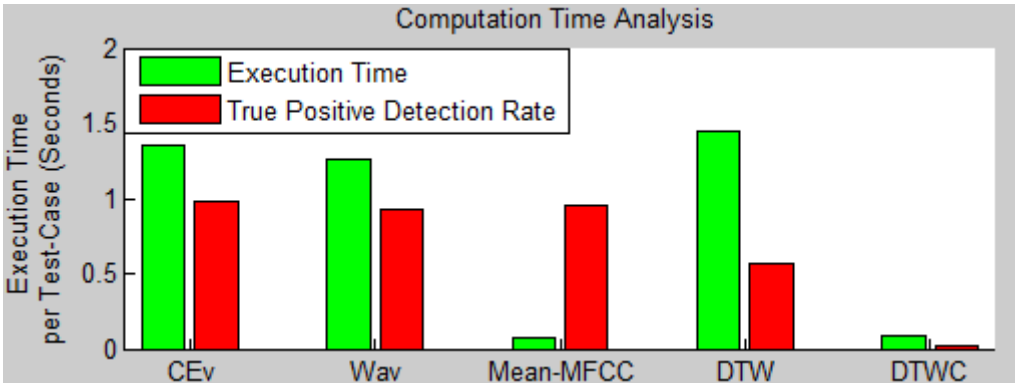

Fig. 3. Keyword detection rate vs. computational cost.

Despite of the KWS performance by the aforementioned approaches, it is also important to analyze the computation time. Fig. 3 provides an indication of execution time for all aforementioned approaches. It is

very clear that the best execution time is achieved by MFFC based approach that is introduced in the current research study. This is because of Euclidean distance deployment for mean values of MFCCs features in the current research rather than DTW which increases the search space as it has been used in the literature. Although, the minimum execution time (i.e. 0.06 sec) is achieved by MFCC features based approach, yet CEv approach with higher execution time (i.e. 1.3 sec) would be preferred because of its superiority in terms of keywords detection rate which is the primary objective. The computation time dramatically decreases by using the constrained DTW; however, it sacrifices a significant amount of keyword detection rate (i.e. 50%) that fails the achievement of the primary objective.

References

[1] Rabiner, L. R., & Juang, B. H. (1993). *Fundamental of Speech Recognition*. New Jersey: PTR Prentice-Hall, 37-51.

[2] Chen, G. (2014). *Low Resource Keyword Spotting.* Department of Electrical and Computer Engineering. Johns Hopkins University. From http://www.clsp.jhu.edu/~guoguo/papers/thesis_proposal.pdf

[3] Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *26(1)*, 43–49.

[4] Zhang, Y., & Glass, J. R. (2009). Unsupervised spoken keyword spotting via segmental DTW on Gaussian posterior grams. *IEEE Proceedings of the Automatic Speech Recognition & Understanding* (ASRU) *Workshop* (pp. 398–403).

[5] Zhang, Y., & Glass, J. R. (2011). An inner-product lower-bound estimate for dynamic time warping. *IEEE Proceedings of the International Conference on Acoustics, Speech and Signal Processing* (pp. 5660–5663).

[6] Chan, C., & Lee, L. (2010). Unsupervised spoken-term detection with spoken queries using segment-based dynamic time warping. *Processing of Inter-Speech*, 693–696.

[7] Fousek, P., & Hermansky, H. (2006). Towards ASR based on hierarchical posterior-based keyword recognition. *IEEE Proceedings of the International Conference on Acoustics, Speech and Signal Processing*: Vol. 1.

[8] Hazen, T. J., Shen, W., & White, C. (2009). Query-by-example spoken term detection using phonetic posterior gram templates. *IEEE Proceedings of the Automatic Speech Recognition & Understanding (ASRU) Workshop* (pp. 421–426).

[9] Parada, C., Sethy, A., & Ramabhadran, B. (2009). Query-by-example spoken term detection for OOV terms. *IEEE Proceedings of the Automatic Speech Recognition & Understanding Workshop* (pp. 404–409).

[10] Wang, H., Lee, T., & Leung, C. (2011). Unsupervised spoken term detection with acoustic segment model. *IEEE Proceedings of the International Conference on Speech Database and Assessments* (pp. 106–111).

[11] Bahi, H., & Benati, N. (2009). A new keyword spotting approach. *IEEE International Conference on Mujltimedia Computing and Systems* (pp. 77-80). Ouarzazate.

[12] MIT. (2003). *Dynamic Time Warping & Search*, *Lecture 9*. Retrieved March 13, 2011, from http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-345-automatic-speech-recognition-spring-2003/lecture-notes/lecture9.pdf

[13] Sharma, P., & Rajpoot, A. K. (2013). Automatic identification of silence, unvoiced and voiced chunks in speech. *Journal of Computer Science & Information Technology*, *3(5)*, 87-96.

[14] Giannakopoulos, T. (2014). A method for silence removal and segmentation of speech signals, implemented in Matlab. Retrieved May 13, 2014, from http://cgi.di.uoa.gr/~tyiannak/Software.html

[15] Khan, W., & Holton, R. (2015). Time warped continuous speech signal matching using Kalman filter. *International Journal of Speech Technology*, *18(1)*, 1381-2416.

[16] Zahorian, S. A., & Hu, H. (2008). A spectral/temporal method for robust fundamental frequency

tracking. *Journal of Acoustic Society of America*, *123(6)*, 4559-4571.

[17] Dave, N. (2013). Feature extraction methods LPC, PLP, and MFCC in speech recognition. *International Journal for Advance Research in Eng. and Tech*, *1(6)*, 1-5.

[18] Foley, B. G. (2012). *A Dempster-Shafer Method for Multi-Sensor Fusion.* MSc thesis, Department of Mathematics and Statistics, Air Force Institute of Technology. Retrieved December 2013, from http://www.dtic.mil/dtic/tr/fulltext/u2/a557749.pdf

[19] Soluade, O. A. (2010). Establishment of confidence threshold for interactive voice response systems using ROC analysis. *Communications of the IIMA*, *10(2)*, 43-57.

[20] Michael, E. (2001). *Top 100 Speeches*. American Rhetoric. Retrieved December 12, 2013, from http://www.americanrhetoric.com/top100speechesall.html

Wasiq Khan was born in Pakistan. He received his BSc degree in math & physics and MSc degree in computer science from COMSATS IIT, Pakistan. At Bradford University (UK), he completed the MSc degree in artificial intelligence. His Ph.D. research at Bradford University is focused on speech processing and template matching while his current research interests include image processing, decision support system, artificial intelligence, machine learning, and speech recognition. Currently, Khan is working as a research associate in artificial intelligence and aero-space sector at University of Central Lancashire, UK. His recent publications cover the time warped speech matching and spoken term detection.

Rob Holton is the head of computer science at Bradford University, UK. He holds a BSc degree in computer science and completed his Ph.D. degree from Queen University, UK. He teaches a number of modules that include programing language theory & semantics, formal methods, and foundations of cryptography. Holton has a number of journal papers, conference papers, book chapters, and research reports. His area of interests includes networks and performance engineering, formal methods, cryptography, and theory and semantics in programing languages.